Markerless Outdoor Human Motion Capture Using Multiple Autonomous Micro Aerial Vehicles

Supplementary Material

Extended Details of The Experiment Results: In section 4.3 of the submitted paper, we discuss the ground truth data collection. Here, we discuss it in details explaining the imperfections.

The IMU data is collected using 17 sensors on the subject's body. These sensors measure data at the rate of 60Hz. The measurements do not have the information about exact time they are recorded. They are sent to the base station using wireless communication. At the time of arrival they are assigned with the frame number (different than our system). Using SIP, we get SMPL parameters for each of these frames. We manually align this sequence to our system by matching a characteristic motion sequence. This way we get unix timestamp for one frame and assuming the frame rate to be 60Hz we get timestamp for every frame. We re-sample the frames which are closest to our image data sequence and compare them for all the error calculations.

In this procedure, we consistently assume that IMU system has a frame rate of 60Hz. However, this is not always the case. The communication delays and failures between the sensors and the base station can cause it to vary. In the video we see that sometimes both the meshes goes out of sync. One possible reason for this is the variable frame rate of both our system and the IMU system. This also affects our quantitative results adversely. Our calculated error is more than what it actually should be.

Ablation study: In Fig 1, we show the advantage of optimizing the camera parameters during human pose estimation. We compare the mean error for each joint in three cases. Case1: we use the camera extrinsics from the online run (i.e. step 1 as mentioned in the paper) and do not optimize them during pose estimation step. Case2: We use the camera extrinsics from the offline run (i.e. step 2 in paper) and do not optimize them during pose estimation step. Case3: we jointly optimize camera extrinsics and pose (our proposed method). For most of the joints, the error in Case2 is lower than Case1 and lowest for all the joints in Case3. Further, we compare the mean of the first order difference of joint positions in all the cases with reference to that of the ref. A lower value closer to the ref implies a smoother and more accurate motion estimate. We can see that the value decreases when going from Case1 to 3, getting closer to ref for all the joints. Since we deal with outdoor, unstructured scenarios we can use only mobile cameras that are autonomously controlled. Getting highly accurate reference (ref) extrinsics of the mobile cameras with minimal on-board computation is extremely difficult. Hence, for our problem, it is important to estimate (and optimize for) both the person's pose and the camera extrinsics, simultaneously.

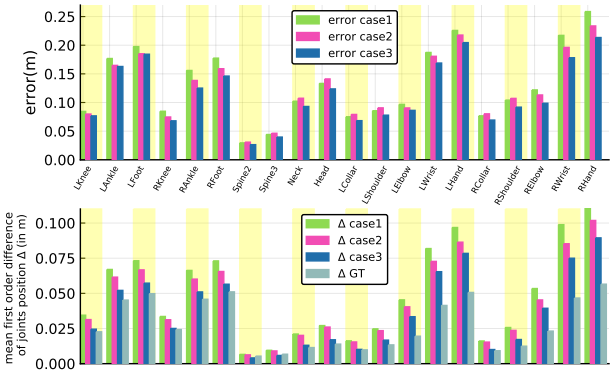

Figure 1: Ablation study of our approach.